「Kabayaki」
エンタープライズ版
for Linux

取扱説明書
Version 1.0.0

第 2 版　2005 年 8 月 23 日

はじめに....1
Kabayaki Enterprise の管理とは....1
コンテンツとインデックス....1

第 1 章 管理画面の構成....3
管理画面の表示....3

第 2 章 サーバー関連メニュー....5
サーバー環境情報....5
検索ログ分析....7
過去 12ヶ月の統計ページ....7
特定年月の統計ページ....13

第 3 章 インデックス関連メニュー....17
インデックス一覧....17
インデックスの追加....19
コンテンツ設定....21
フィルタ設定....23
Web スパイダ....25
基本設定....25
拡張設定....25
チューニング設定....27
ログ一覧....32

第 4 章 検索および検索結果画面....35
検索画面の表示....35
検索方法....36
検索式....36

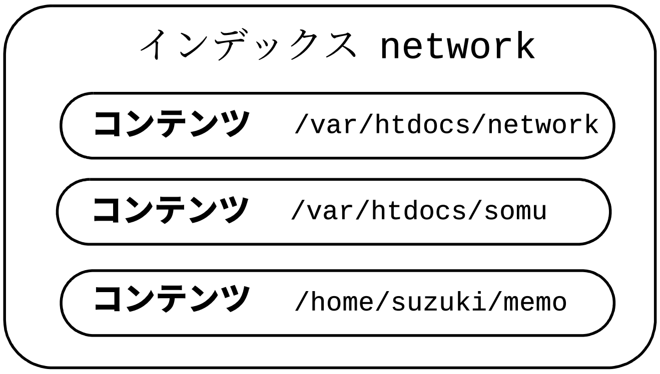
付録 A Kabayaki Enterprise 文書フィルタとプロパティ検索詳細....41
文書フィルタ....41
対応文書....41
プロパティ検索....42
OLE オブジェクト検索....42

はじめに

Kabayaki Enterprise の管理とは

コンテンツとインデックス

「Kabayaki」エンタープライズ版 (以後、Kabayaki Enterprise と表記) は、コンテンツとインデックスという枠組みを使用して文書を検索します。コンテンツとインデックスとは次のような関係になります。



上の図では、Kabayaki Enterprise を運用するシステム上のファイルシステムでは、/var/htdocs/network、/var/htdocs/somu、/home/suzuki/memo という別々の場所で管理している社内ネットワーク関連文書を、network という１つのインデックスにまとめて管理しています。

第 1 章　管理画面の構成

管理画面の表示

Kabayaki Enterprise の管理には、Microsoft Internet Explorer や Netscape Communications Netscape Navigator などの Web ブラウザを使用します。管理画面を表示させるには、Web ブラウザへ次のように URL を入力します。

http:// *ホスト名* /kabayaki/cgi-bin/admin/rc.cgi

「*ホスト名*」の部分には、Kabayaki Enterprise をインストールしたコンピュータの名前を入力します。

たとえば、Kabayaki Enterprise をインストールしたホストが search.timedia.co.jp ならば、次の URL になります。

http://search.timedia.co.jp/kabayaki/cgi-bin/admin/rc.cgi

Kabayaki Enterprise が正しくインストールされていると、管理画面が表示されます。

Kabayaki Enterprise **をインストールした状態での画面イメージ** :

管理画面左側の一覧をメインメニューと呼びます。メインメニューは、サーバー関連メニューとインデックス関連メニューから構成され、オプションパックである kabayaki-dbspider パッケージがインストールされている場合は、さらに一覧が追加されます。

サーバー関連メニューでは、Kabayaki Enterprise をインストールしたホスト全体に関わる情報の設定や表示の種類を選べます。インデックス関連メニューでは、検索に必要なインデックスの情報を設定する機能を選べます。メインメニューは、機能を選んで画面が切り替わっても、常に表示されています。

Kabayaki Enterprise をインストールして、最初に Kabayaki Enterprise 管理画面を表示させたときは、サーバー関連メニューの機能とインデックス関連メニューのインデックス一覧しか選ぶことができません。その他の一覧や設定を選ぶには、後述する手順でインデックスを作成する必要があります。

各設定画面に共通で表示されるものには、さらに次のものがあります。

「?」(ヘルプ)ボタン

ページ右上に表示されます。クリックすると、各設定画面のオンラインヘルプが表示されます。

「インデックス選択」

インデックス関連メニューを選択したときのみ表示されます。ページの右上の方に表示されているインデックス名の右の矢印をクリックすると、インデックス名の一覧がメニュー表示されます。メニューからインデックスを選択すると、表示および操作の対象となるインデックスが変更されます。

第 2 章　サーバー関連メニュー

この章では、「サーバー関連メニュー」に分類されている機能のうち、辞書関連のメニューを除いたメニュー (「サーバー環境情報」と「検索ログ分析」) について説明します。

類語関連および形態素関連については、『Kabayaki エンタープライズ版 for Linux　辞書管理ツール説明書』を参照してください。

サーバー環境情報

Kabayaki Enterprise をインストールしたホストに関する情報を表示します。また、検索および検索結果の画面のデザインを変更できます。管理画面の左側に表示されるメニューの「サーバー環境情報」ボタンをクリックすると、この「サーバー環境情報」画面が表示されます。

画面に表示されている情報は以下の通りです。

ホスト名

Kabayaki Enterprise が動作しているホストの名前が表示されます。環境変数 SERVER_NAME を参照しています。SERVER_NAME が設定されていないと、「不明」と表示されます。

ホスト IP アドレス

Kabayaki Enterprise が動作しているホストの IP アドレスが表示されます。これは、環境変数 SERVER_ADDR を参照しています。SERVER_ADDR が設定されていない場合は、「不明」と表示されます。

実行ユーザー

Kabayaki Enterprise 管理画面を実行しているプロセスの実行ユーザ名が表示されます。

検索ページテンプレートタイプ

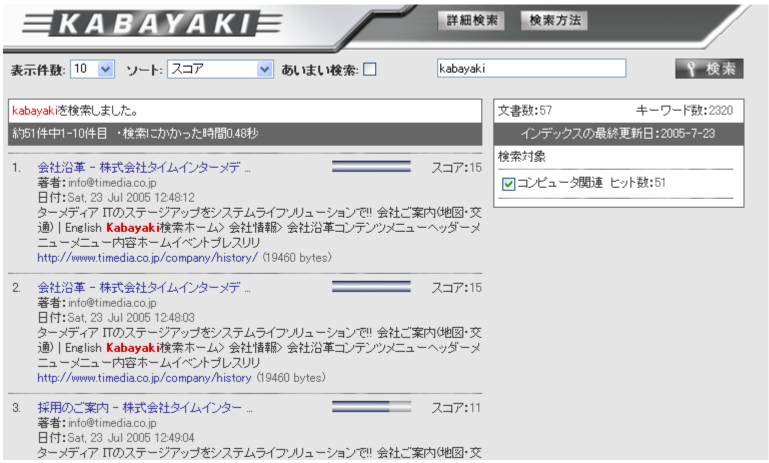
検索と検索結果の画面の外観を選択できます。enterprise はインストールした直後と同じ画面になります。

「保存」ボタン

「保存」ボタンをクリックすると、検索ページテンプレートタイプの設定が保存されます。

検索テンプレートが使用された検索結果画面

enterprise テンプレート：



検索ログ分析

「検索ログ分析」では、ユーザーがサイト上で実行した検索について様々な角度から分析することができます。

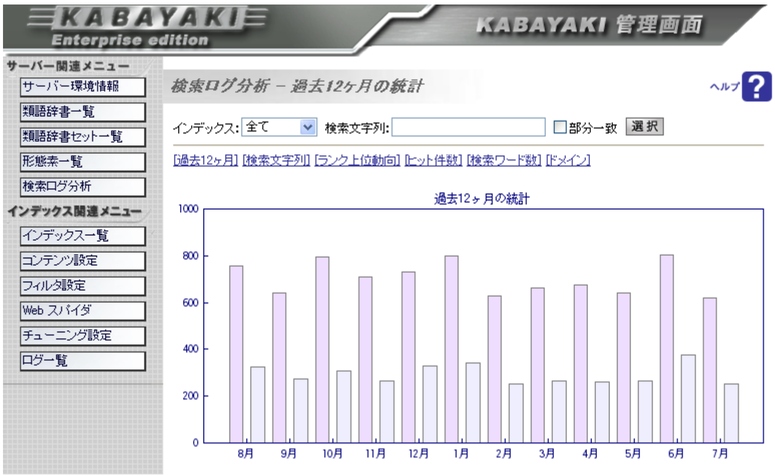
過去 12ヶ月の統計ページと、特定年月の統計（YYYY 年 MM 月の統計）ページから構成され、それぞれのページで検索回数の時系列的な動きや、各種ランキング情報が提供されます。

なお、本ページで表示する情報は、通常夜間・早朝に実行される集計処理の際に更新されます。

過去 12ヶ月の統計ページ

過去 12ヶ月の統計結果を表示します。画面左側のナビゲーションの「検索ログ分析」をクリックすると、最初に表示されるページとなります。

ページ内のタイトル部分には、「検索ログ分析 - 過去 12ヶ月の統計」と表示され、ページの上から順に、絞り込み検索欄、ナビゲーション欄、統計結果表示欄の順で構成されます。



絞り込み検索欄

絞り込み検索欄では、様々な条件により統計データを絞り込み可能です。絞り込み条件などの各項目の説明は以下の通りです。

[インデックス] 選択メニュー

メニューからインデックスを選択して統計結果を絞り込みます。デフォルトは「全て」です。

なお、本ページのインデックスメニューは、他の統計情報と同様、通常夜間・早朝に実行される集計の際に更新されます。

[検索文字列] 入力ボックス

テキスト入力された検索文字列によって統計結果を絞り込みます。デフォルトは空欄（絞込なし）です。

[部分一致] チェックボックス

チェックの有 / 無により、検索文字列の部分一致検索の有効 / 無効を設定。デフォルトはチェックなし (部分一致検索無効）です。

[選択] ボタン

上記条件による絞り込みを実行します。

ナビゲーション欄

ナビゲーション欄には各統計へのリンクがあり、過去 12ヶ月の各統計部分へ即座に移動可能です。

各リンクの説明は以下の通りです。

[過去 12ヶ月]

過去 12ヶ月の統計（グラフ / 表）へのリンクです。

[検索文字列]

検索文字列 TOP20（表）へのリンクです。

- 完全一致で検索文字列の絞り込みを行った場合、本リンクは表示されません。

[ランク上位動向]

ランキング上位の動向（グラフ）へのリンクです。

- 完全一致で検索文字列の絞り込みを行った場合、本リンク表示されません。

[ヒット件数]

ヒット件数（グラフ / 表）へのリンクです。

[検索ワード数]

検索ワード数（グラフ / 表）へのリンクです。

- 完全一致で検索文字列の絞り込みを行った場合、本リンクは表示されません。

[ドメイン]

ドメイン （グラフ / 表）へのリンクです。

統計結果表示欄

様々な統計を表やグラフとして表示します。

[過去 12ヶ月の統計] グラフ

検索回数 / 検索文字列数の過去 12ヶ月の動向を棒グラフで表示します。

[過去 12ヶ月の統計] 表

検索回数 / 検索文字列数の過去 12ヶ月の統計表を表示します。

過去12ヶ月の統計				
年月	一日あたりの平均		月合計	
	検索回数	検索文字列数	検索回数	検索文字列数
2004年8月	24.4	10.4	755	323
2004年9月	21.3	9.2	640	275
2004年10月	25.6	10.0	793	309
2004年11月	23.6	8.8	709	265
2004年12月	23.5	10.5	730	327
2005年1月	25.8	11.1	800	343
2005年2月	22.4	9.0	628	252
2005年3月	21.4	8.5	664	265
2005年4月	22.5	8.6	675	259
2005年5月	20.6	8.6	639	266
2005年6月	26.8	12.5	803	374
2005年7月	22.0	9.1	617	255
総合計			8453	2401

→ 過去12ヶ月の統計のCSV出力

検索回数 / 検索文字列数それぞれについて、1 日当たりの平均と月の総計を表示します。

「年月」列の各データ部分（yyyy 年 mm 月）をクリックすると、特定年月の統計ページへ移動します。

[過去 12ヶ月の統計の CSV 出力] リンク

検索回数 / 検索文字列数の統計データを CSV 形式で出力します。

[検索文字列 TOP20] 表

検索回数の多かった検索文字列の上位 20 件をランキング表示します。

検索文字列 TOP20			
順位	合計		検索文字列
1	2048	24.23%	蒲焼き
2	273	3.23%	検索
3	226	2.67%	namazu
4	215	2.54%	kabayaki
5	116	1.37%	test
6	81	0.96%	蒲焼
7	80	0.95%	Kabayaki
8	69	0.82%	ダウン ロード
9	63	0.75%	あ
10	56	0.66%	KABAYAKI
11	54	0.64%	インストール
12	54	0.64%	テスト
13	53	0.63%	価格
14	52	0.62%	やき
15	46	0.54%	まず
16	45	0.53%	pdf
17	40	0.47%	サイト
18	36	0.43%	PDF
19	36	0.43%	kaba
20	36	0.43%	シソーラス

→ 検索文字列 TOP20のCSV出力

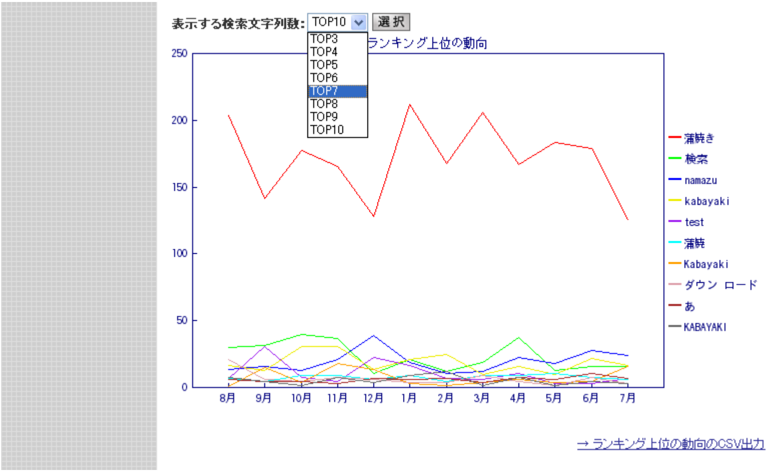
「検索文字列」列の各データ部分をクリックすると検索文字列で絞り込みを行います。

[検索文字列 TOP20 の CSV 出力] リンク

検索文字列 TOP20 のランキングデータを CSV 形式で出力します。

[ランキング上位の動向] グラフ

検索回数ランキング上位の検索文字列について、検索回数の動向を折れ線グラフで表示します。



- グラフ左上のメニューにより、グラフで表示する検索文字列の数を選択可能です。
- 完全一致で検索文字列の絞り込みを行った場合、本グラフは表示されません。

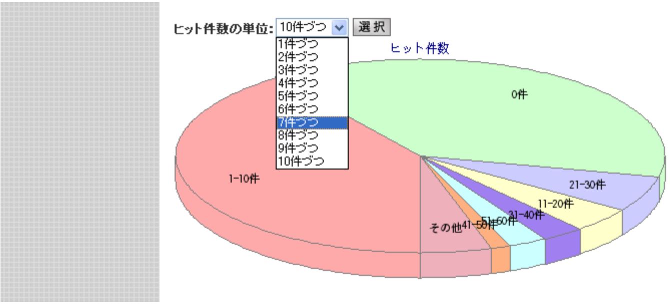
[ランキング上位の動向の CSV 出力] リンク

ランキング上位の動向の統計データを CSV 形式で出力します。

- 完全一致で検索文字列の絞り込みを行った場合、本リンクは表示されません。

[ヒット件数 TOP20] グラフ

ヒット件数（検索条件に一致したページ数）について、上位のものを円グラフで表示します。



- グラフ左上のメニューにより、ヒット件数の単位を選択可能です。
- 総件数が 0 件の場合、本グラフは表示されません。

[ヒット件数 TOP20] 表

ヒット件数（検索条件に一致したページ数）の上位 20 件をランキング表示します。

ヒット件数 TOP20

順位	合計		ヒット件数
1	3375	39.93%	1-10件
2	3262	38.59%	0件
3	509	6.02%	21-30件
4	348	4.12%	11-20件
5	233	2.76%	31-40件
6	218	2.58%	51-60件
7	112	1.32%	41-50件
8	91	1.08%	61-70件
9	53	0.63%	121-130件
10	53	0.63%	371-380件
11	34	0.40%	111-120件
12	32	0.38%	91-100件
13	30	0.35%	101-110件
14	27	0.32%	251-260件
15	22	0.26%	271-280件
16	15	0.18%	81-90件
17	9	0.11%	71-80件
18	6	0.07%	211-220件
19	4	0.05%	231-240件
20	4	0.05%	331-340件

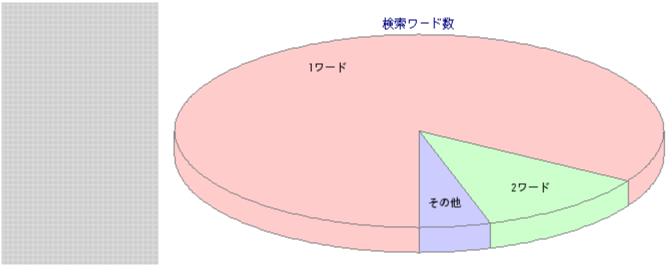
→ヒット件数 TOP20のCSV出力

[**ヒット件数** TOP20 **の** CSV **出力**] **リンク**

ヒット件数（検索条件に一致したページ数）TOP20 のランキングデータを CSV 形式で出力します。

[**検索ワード数** TOP20] **グラフ**

検索文字列を構成するワード数について、上位のものを円グラフで表示します。



- 完全一致で検索文字列の絞り込みを行った場合、本グラフは表示されません。
- 総件数が 0 件の場合、本グラフは表示されません。

[**検索ワード数** TOP20] **表**

検索文字列を構成するワード数の上位 20 件をランキング表示します。

検索ワード数 TOP20			
順位	合計		検索ワード数
1	7076	83.71%	1ワード
2	976	11.55%	2ワード
3	249	2.95%	3ワード
4	73	0.86%	4ワード
5	29	0.34%	5ワード
6	14	0.17%	7ワード
7	10	0.12%	16ワード
8	8	0.09%	32ワード
9	7	0.08%	6ワード
10	6	0.07%	8ワード
11	2	0.02%	15ワード
12	2	0.02%	17ワード
13	1	0.01%	10ワード

→ 検索ワード数 TOP20のCSV出力

- 完全一致で検索文字列の絞り込みを行った場合、本表は表示されません。

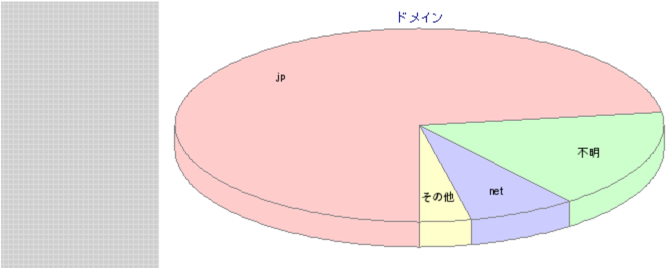
[検索ワード数 TOP20 の CSV 出力] リンク

検索ワード数 TOP20 のランキングデータを CSV 形式で出力します。

- 完全一致で検索文字列の絞り込みを行った場合、本表は表示されません。

[ドメイン TOP20] グラフ

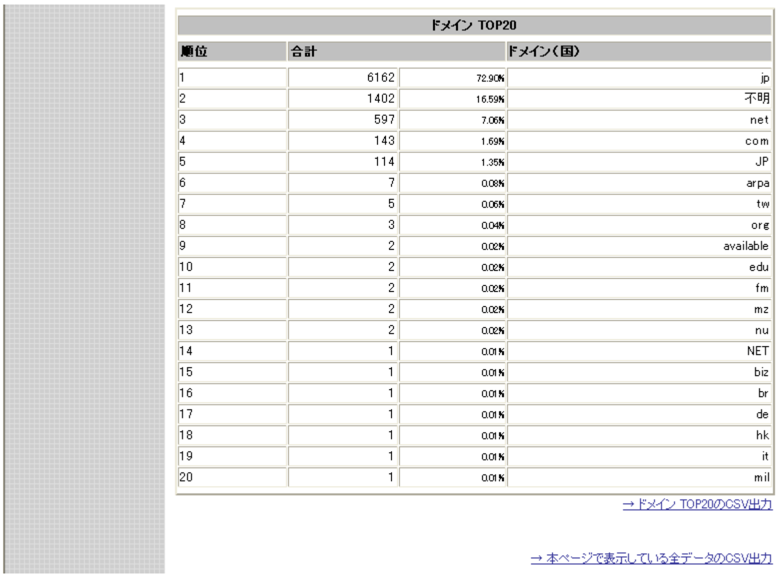
アクセス元のトップレベルドメイン (jp, com, net 等) の割合を円グラフで表示します。



- 総件数が 0 件の場合、本グラフは表示されません。

[ドメイン TOP20] 表

アクセス元のトップレベルドメイン (jp, com, net 等) の上位 20 件をランキング表示します。



ドメイン TOP20			
順位	合計		ドメイン（国）
1	6162	72.90%	jp
2	1402	16.59%	不明
3	597	7.06%	net
4	143	1.69%	com
5	114	1.35%	JP
6	7	0.08%	arpa
7	5	0.06%	tw
8	3	0.04%	org
9	2	0.02%	available
10	2	0.02%	edu
11	2	0.02%	fm
12	2	0.02%	mz
13	2	0.02%	nu
14	1	0.01%	NET
15	1	0.01%	biz
16	1	0.01%	br
17	1	0.01%	de
18	1	0.01%	hk
19	1	0.01%	it
20	1	0.01%	mil

[ドメイン TOP20 の CSV 出力] リンク

ドメイン TOP20 のランキングデータを CSV 形式で出力します。

[本ページで表示している全データの CSV 出力] リンク

本ページで表示している全データを CSV 形式で出力します。

特定年月の統計ページ

特定年月の統計結果が表示されます。

本ページは、[過去 12ヶ月の統計] 表の「年月」列の各データ部をクリックすると表示され、ページ内のタイトル部分には、「検索ログ分析 - YYYY 年 MM 月の統計」と表示されます。YYYY と MM の部分にはそれぞれ実際の年と月が入ります。

2005年1月の統計		
検索回数		800
検索文字列数		343
	平均	最大
一時間あたりの検索回数	1.1	29
一時間あたりの検索文字列数	0.5	17
一日あたりの検索回数	25.8	101
一日あたりの検索文字列数	11.1	50

ページ構成は過去 12ヶ月の統計と同じで、ページの上から順に、絞り込み検索欄、ナビゲーション欄、統計結果表示欄となります。

絞り込み検索欄

絞り込み検索欄では、様々な条件により特定年月の統計データを絞り込みます。

絞り込み条件は過去 12ヶ月の統計のものと同じです。

ナビゲーション欄

ナビゲーション欄のリンクは、過去 12ヶ月のものと同じものに加え、[日ごと] と [時間ごと] が追加されています。

また [過去 12ヶ月] のみ他のリンクとパイプ（|）で分けられ、他のリンクとは異なり別ページへのリンクとなります。

それぞれの説明は以下の通りです。

[過去 12ヶ月]

過去 12ヶ月のページへのリンクです。唯一別ページへのリンクとなります。

[日ごと]

日ごとの統計（グラフ / 表）へのリンクです。

[時間ごと]

時間ごとの統計（グラフ / 表）へのリンクです。

統計結果表示欄

様々な統計を表やグラフとして表示します。
過去 12ヶ月の統計ページと大部分が同じです。以下では過去 12ヶ月の統計ページと異なる部分について説明します。

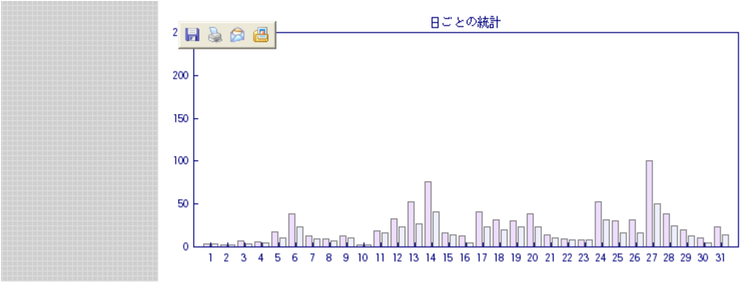
[YYYY 年 MM 月の統計] 表

特定年月の統計のサマリーを表示します。

検索回数 / 検索文字列数の総計、一時間あたりの検索回数 (平均 / 最大)、一時間あたりの検索文字列数（平均 / 最大)、一日あたりの検索文回数（平均 / 最大)、一日あたり検索文字列数（平均 / 最大）を表示します。

[日ごとの統計] グラフ

日ごとの検索回数 / 検索文字列数の統計を棒グラフで表示します。



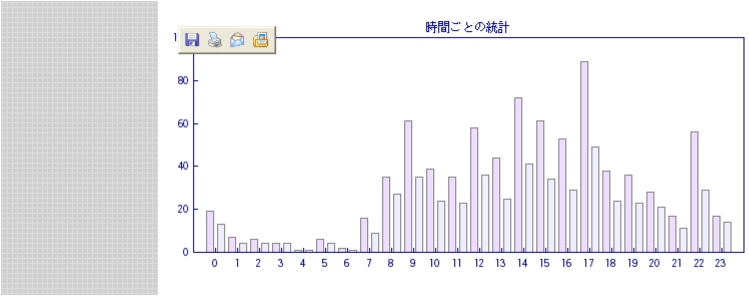
[日ごとの統計] 表

日ごとの検索回数 / 検索文字列数の統計表を表示します。

日ごとの統計				
日	検索回数		検索文字列数	
1日	4	0.50%	4	1.17%
2日	2	0.25%	2	0.58%
3日	7	0.88%	4	1.17%
4日	6	0.75%	5	1.46%
5日	17	2.12%	11	3.21%
6日	38	4.75%	23	6.71%
7日	13	1.62%	9	2.62%
8日	9	1.12%	7	2.04%
9日	13	1.62%	10	2.92%
10日	2	0.25%	2	0.58%
11日	19	2.38%	16	4.66%
12日	33	4.12%	23	6.71%
13日	52	6.50%	27	7.87%
14日	76	9.50%	41	11.95%
15日	16	2.00%	14	4.08%
16日	13	1.62%	5	1.46%
17日	41	5.12%	23	6.71%
18日	32	4.00%	20	5.83%
19日	30	3.75%	23	6.71%
20日	39	4.88%	23	6.71%
21日	14	1.75%	10	2.92%
22日	9	1.12%	8	2.33%

[時間ごとの統計] グラフ

時間ごとの検索回数 / 検索文字列数の統計を棒グラフで表示します。



[時間ごとの統計] 表

時間ごとの検索回数 / 検索文字列数の統計表を表示します。

時間ごとの統計

時	検索回数			検索文字列数		
	平均	合計		平均	合計	
0時	0.6	19	2.38%	0.4	13	3.79%
1時	0.2	7	0.88%	0.1	4	1.17%
2時	0.2	6	0.75%	0.1	4	1.17%
3時	0.1	4	0.50%	0.1	4	1.17%
4時	0.0	1	0.12%	0.0	1	0.29%
5時	0.2	6	0.75%	0.1	4	1.17%
6時	0.1	2	0.25%	0.0	1	0.29%
7時	0.5	16	2.00%	0.3	9	2.62%
8時	1.1	35	4.38%	0.9	27	7.87%
9時	2.0	61	7.62%	1.1	35	10.20%
10時	1.3	39	4.88%	0.8	24	7.00%
11時	1.1	35	4.38%	0.7	23	6.71%
12時	1.9	58	7.25%	1.2	36	10.50%
13時	1.4	44	5.50%	0.8	25	7.29%
14時	2.3	72	9.00%	1.3	41	11.95%
15時	2.0	61	7.62%	1.1	34	9.91%
16時	1.7	53	6.62%	0.9	29	8.45%

注意制限事項

- データ件数が多い場合、表示に時間のかかることがあります。
- 一日のうちで同じページを同じ絞込条件で閲覧する場合、前回のデータを内部的に記憶しているため表示が早くなります。
- 検索文字列等では、全体に対する割合をパーセンテージ（%）で少数点第 2 位まで表示していますが、これらを合計しても 100% にならない場合があります。
- 検索回数 / 検索文字列数はインデックスごとにカウントされます。例えば 2 つのインデックスに対して 1 回検索を行った場合、検索回数は 2 となります。
- 部分一致が無効の場合大文字小文字は区別されますが、部分一致が有効の場合大文字小文字は区別されません。また全角英数字は半角英数字へ自動的に変換されます。
- 部分一致検索を有効にした場合、絞込に時間がかかる場合があります。

第３章　インデックス関連メニュー

インデックス一覧

管理画面の左側に表示されるメニューの「インデックス一覧」ボタンをクリックすると、「インデックス一覧」画面が表示され、登録されているインデックスが一覧表示されます。

Kabayaki Enterprise をインストールした直後など、登録されているインデックスが１つも存在しないときは、

　新しくインデックスを作成してください。

と表示されます。なお、インデックスについては、『はじめに』の「Kabayaki Enterprise の管理とは コンテンツとインデックス」を参照してください。

「インデックス追加」ボタン

入力されている内容で、Kabayaki Enterprise のインデックスを新規に登録します。

インデックス一覧の表示は､｢内部名｣のアルファベット順です。
一覧で表示される項目は次の通りです。

表示名

表示名は、他の Kabayaki Enterprise 管理画面や検索結果画面でも表示されます。クリックすると、そのインデックスの「コンテンツ設定」画面が表示されます。

内部名

内部名は、Kabayaki Enterprise が利用する名前です。表示されている内部名をクリックすると、そのインデックスの「コンテンツ設定」画面が表示されます。

コンテンツ種別

インデックスのコンテンツの種類が表示されます。ここに表示される内容は、インデックス作成時の指定および「コンテンツ設定」画面での設定に応じて決まります。

コンテンツが未設定のインデックスでは「-」が表示され、コンテンツにローカルパス上のファイルが指定されていると「ファイル」、Web コンテンツに http:// で始まるリモートパスが指定されていると「Web」と表示されます。両方が指定されているときは「ファイル /Web」という表示になります。

状態

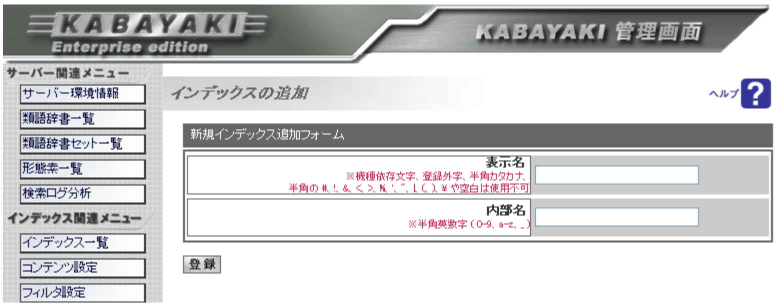
インデクシング中には、ここに「処理中」と表示されます。それ以外のときは「-」が表示されます。

操作

削除ボタンを押すと、そのインデックスを削除するための画面に移動します。

インデックスの追加

「インデックス一覧」画面の「インデックス追加」ボタンをクリックすると、「新規インデックス追加フォーム」が表示されます。



項目は次の通りです。

表示名

Kabayaki Enterprise 管理画面や検索結果画面で表示されるインデックス名を入力します。機種依存文字や登録外字、半角カタカナは使用できません。また、半角の # ! & < > % ' " | ( ) ¥ や空白文字も指定できません。

内部名

インデックス識別のために Kabayaki Enterprise が利用する名前を入力します。半角小文字の英数字とアンダースコア (_) のみが入力できます。

0123456789 abcdefghijklmnopqrstuvwxyz _

内部名、表示名ともに、入力できる文字数の制限を超えて入力することはできません。

インデックスの数は 64 個まで作成・検索可能です。なお、インデックスの内部名の長さや、Web サーバー、Web ブラウザによって、検索可能なインデックスの数は 64 個よりも少なくなることがあります。インデックスの内部名は検索時の GET パラメータとして利用されるため、作成するインデックスの数が多くなる場合は、なるべく短い名前にすることをお勧めします。

「登録」ボタン

入力されている内容で、Kabayaki Enterprise のインデックスを新規に登録します。

注意

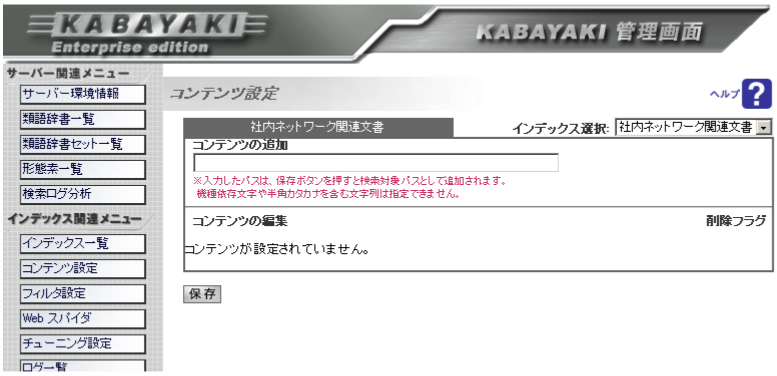
インデクシングが行なわれている最中に、インデックス追加や削除を実行しないでください。このような操作をしますと、以後、インデクシングや検索が正しく動作しなくなることがあります。

インデックスの数は 64 個まで作成・検索可能です。なお、インデックスの内部名の長さや、Web サーバー、Web ブラウザによって、検索可能なインデックスの数は 64 個よりも少なくなることがあります。インデックスの内部名は検索時の GET パラメータとして利用されるため、作成するインデックスの数が多くなる場合は、なるべく短い名前にすることをお勧めします。

コンテンツ設定

インデックスに設定されるコンテンツ (検索対象の文書がある場所) を、追加・設定・削除します。コンテンツについては、『はじめに』の「Kabayaki Enterprise の管理とは コンテンツとインデックス」を参照してください。

管理画面の左側に表示されるメニューの「コンテンツ設定」ボタンをクリックするか、「インデックス一覧」画面で一覧表示されている表示名または内部名のリンクをクリックすると、この「コンテンツ設定」画面が表示されます。



画面に表示されている情報は以下の通りです。

コンテンツの追加

入力フィールドに、インデクシングを実行したい (検索対象にしたい) 文書が含まれるディレクトリのパスを、絶対パスで入力します。検索の対象となるファイルは、指定されたディレクトリ以下の全てのファイルとなります。

このフィールドに機種依存文字や半角カタカナを含む文字列を指定することはできませんので、注意してください。

ローカルパスの他、http:// で始まるリモートパスを指定することも可能です。

コンテンツの編集

コンテンツの一覧が表示されます。インデックスが作成された直後の時点では、設定されているコンテンツは存在しません。

コンテンツの追加フィールドには、検索の結果に表示される文書の場所を指定します。コンテンツ (検索対象) としてローカルパスが指定された直後は、「閲覧時の URL」欄には「http:// ローカルパス /」のようにローカルパスがそのまま入るため、検索を実行したユーザーの Web ブラウザから文書を参照可能な URL にするための変更が必要になることがあります。たとえば、コンテンツが存在するディレクトリが

```
/var/www/html
```

に設定されている Web サーバー search.timedia.co.jp の場合は、次のように「閲覧時の URL」を編集します。

http://var/www/html/yamada → http://search.timedia.co.jp/yamada

http://var/www/html/yamada_syanaihi → http://search.timedia.co.jp/yamada_syanaihi

http://usr/local/apache/htdocs/yamada → http://search.timedia.co.jp/yamada

「削除フラグ」チェックボックス

コンテンツを削除するには、「コンテンツの編集」の一覧の、削除したいコンテンツの右側に表示されている「削除フラグ」チェックボックスをチェックしてから、「保存」ボタンをクリックします。

文書の格納されている場所によっては、Web ブラウザで検索結果を正しく表示させるために、Web サーバーの設定ファイルの編集が必要になることがあります。

Apache HTTP Server では、標準の設定ファイルである httpd.conf を編集します。たとえば、以下のような行を httpd.conf に追加します。

```
Alias    /doc/network      "/usr/local/doc/network"
```

コンテンツの追加で、異なるディレクトリパスを設定するたびに、上記のような Alias ディレクティブを追加していきます。

Alias を追記したら、Apache HTTP Server を再起動して、httpd.conf の変更を反映させせます。

```
# apachectl graceful
```

または

```
# httpd restart
```

フィルタ設定

ファイルの拡張子で示されるファイル形式やサブディレクトリを指定して、検索の対象にするコンテンツを選別 (フィルタリング) することができます。フィルタ設定を上手に利用することによって、無駄なファイルのインデクシングを回避し、インデクシング時間やホストの資源を節約することができます。設定項目の優先順位が低い順に並べると、次の通りになります。

優先順位　低　←――――――――→　高　優先順位

検索対象ファイル ＜ 検索対象外ファイル拡張子 ＜ 検索対象外パス

管理画面の左側に表示されるメニューの「フィルタ設定」ボタンをクリックすると、この「コンテンツ設定」画面が表示されます。

検索対象ファイル

検索の対象にしたいファイル名の拡張子を選び、チェックボックスをチェックします。
「その他のファイル」をチェックすると、項目にない全てのファイルを検索の対象とします。

- HTML ファイルは拡張子が以下のものを対象とします。
  html, htm, phtml, shtml, html.(英数二文字)
- Mail/News, man 形式は以下のものを対象とします。
  数字のみ、または、文字 + 数字
- 一太郎は拡張子が以下のものを対象とします。
  jsw, jaw, jbw, jfw, jtd

注意事項：

「その他のファイル」をチェックした場合でも、以下のファイルは検索対象となりません。

- アーカイブファイル (*.tar, *.tgz, *.lzh, *.zip)
- Windows システムファイル (*.exe, *.dll)
- Microsoft Visio ファイル (*.vsd)
- Microsoft Project ファイル (*.mpp)
- Microsoft Access ファイル (*.mdb)
- メディアファイル (*.wav, *.wmv, *.wmz, *.swf)
- 画像ファイル (*.psd, *.ai, *.gif, *.png, *.jpg,*.jpeg, *.dib, *.bmp, *.tif, *.tiff)
- # で始まるファイル

検索対象外ファイル

検索の対象にしたくないファイル名の拡張子をテキストフィールド内に記述します。検索対象ファイルの設定でその他のファイルを検索対象とした際、検索の対象外にしたいファイルの設定に役立ちます。

例）
*.sit
*.c

検索対象外パス

検索の対象にしないディレクトリのパスを指定します。「コンテンツ設定」で指定されているパスの下に存在するが、検索の対象には含めたくないサブディレクトリを指定します。

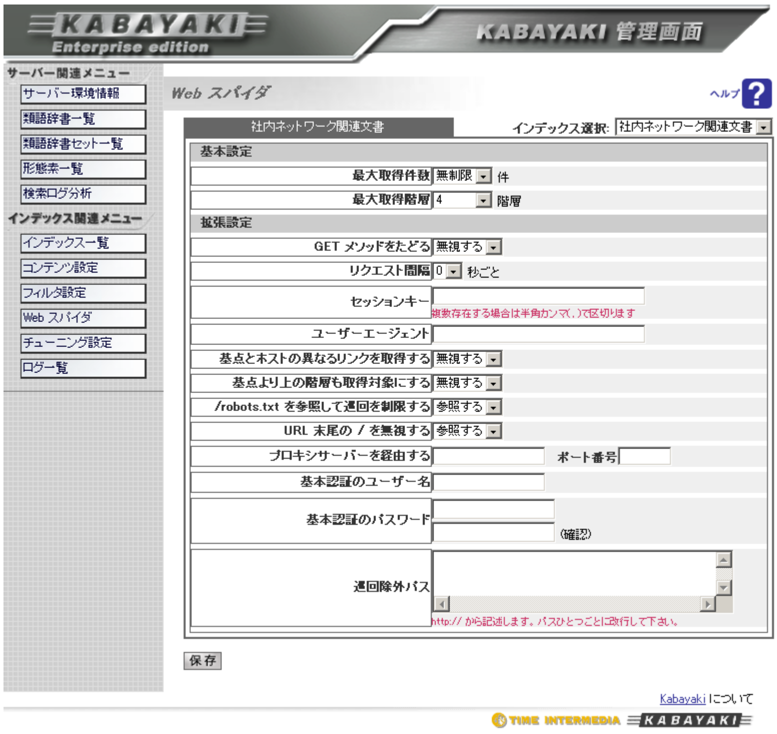
このフィールドに機種依存文字や半角カタカナを含む文字列を指定することはできませんので、ご注意ください。また、ここで指定できるのはローカルパスのみです。http:// で始まるリモートパスは、「Web スパイダ」の「巡回除外パス」で設定してください。

「保存」ボタン

「保存」ボタンをクリックすると、入力されているフィルタ設定を保存します。

Web スパイダ

Web スパイダは、数多くの Web サイトを巡回し、文章や画像などのコンテンツを収集するアプリケーションです。Web スパイダは、いろいろな Web サイトの仕組みに対応し、取り込んだコンテンツを Kabayaki Enterprise で検索できるようにするための細かな設定が可能になっています。



基本設定

最大取得件数

収集するコンテンツの数を制限します。初期設定は「無制限」で、収集するコンテンツの数は制限されていません。

最大取得階層

探索するハイパーリンクの階層の数を制限します。コンテンツ設定で入力した URL の直下の下層から数えた階層が探索の対象となります。初期設定は「4」です。

拡張設定

GET メソッドをたどる

URL に表れる ? より右側の & で区切られた = をはさんだキーと値の組み合わせ (クエリー) を、URL の一部とみなすかどうかを設定します。初期値は「無視する」ですので、クエリーを URL の一部とはみなさずに探索します。

リクエスト間隔

Web スパイダからサーバーへコンテンツの取り出しを要求する間隔を秒数で指定します。

セッションキー
セッションキーを取り除いたものを URL とするかどうかを設定します。初期値は空白です。セッションキーにあたる文字列を入力すると、それを取り除いたものを URL と見なして探索します。

ユーザーエージェント
Web サーバーへ送信するユーザーエージェント情報を文字列で設定します。初期値は空白です。

基点とホストの異なるリンクを取得する
異なる Web サーバーのコンテンツも探索するかどうかを設定します。初期設定は「無視する」で、基点となる URL から Web サーバーを越えた探索をしません。

基点より上の階層も取得対象にする
コンテンツ設定で入力した URL を遡って探索するかどうかを設定します。初期設定は「無視する」で、ハイパーリンクや CGI で URL が遡ったコンテンツは探索しません。

/robots.txt を参照して巡回を制限する
robots.txt の内容に従って巡回を制限するかどうかを設定します。初期設定は「参照する」で、robots.txt の内容を遵守して探索します。

URL 末尾の / を無視する
一部のサイトではコンテンツの URL 末尾に「/」(スラッシュ) がついていることがあります。これを取り除いてインデクシングの対象とします。

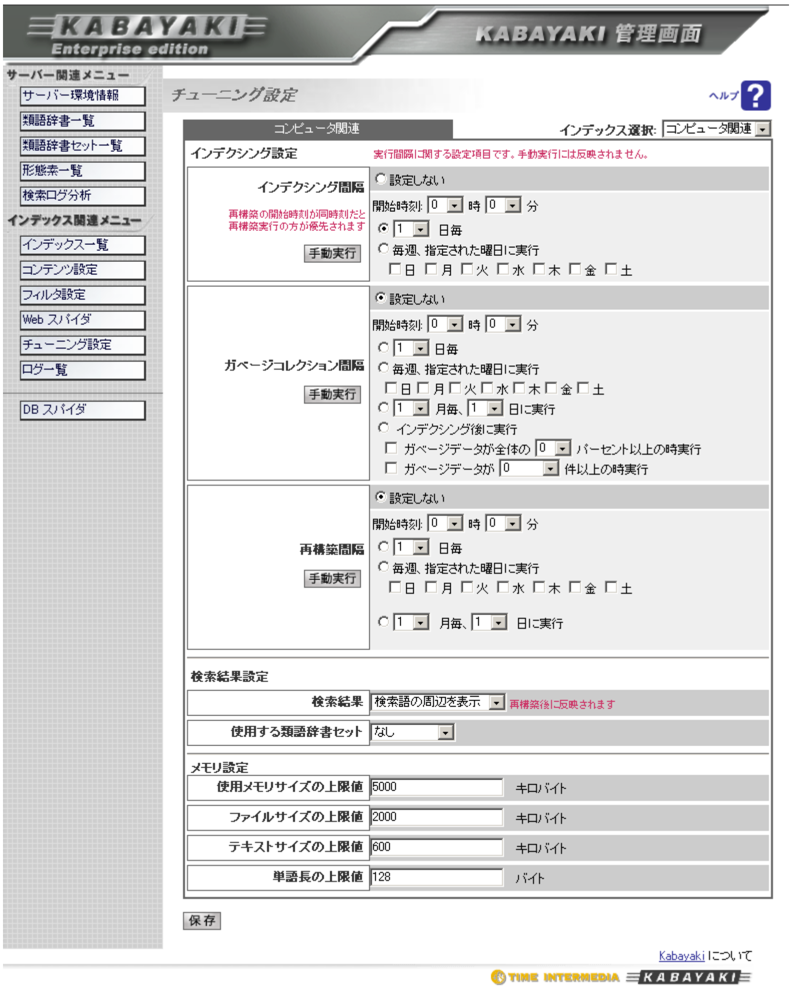
基本認証のユーザー名および基本認証のパスワード
基本認証(BASIC 認証) のあるコンテンツを閲覧する時に必要なユーザー名とパスワードを設定します。初期値は空白で、基本認証の必要なコンテンツは探索しません。

プロキシサーバーを経由する
プロキシ (HTTP PROXY) を経由しないと Web コンテンツを取得できないネットワーク環境にいる場合、経由させるプロキシサーバーのホスト名と使用するポート番号を指定します。

巡回除外パス
Web コンテンツを取得しなくてもよい URL を指定します。除外したい URL が複数あるときは、1 つの URL につき 1 行ずつ入力します。

チューニング設定

「チューニング設定」画面では、日々変化するコンテンツのインデクシングを効率よく管理するための、インデクシング実行の間隔や時刻の設定、メモリ設定等の設定変更機能を提供しています。



インデクシング間隔

インデクシングを実行する間隔、開始時刻、曜日を設定します。インデックス毎に異なった時刻を設定できます。インデクシングの処理対象となるのは、新規追加または更新されたファイルです。

実行間隔を、「設定しない」、「*n* 日毎」、「毎週、指定された曜日に実行」から選択します。

- 「*n* 日毎」、「毎週、指定された曜日に実行」のどちらかを選択すると、「開始時刻 :」で選択した時刻に処理が開始されます。
- 「*n* 日毎」のラジオボタンを選択したときは、何日毎に実行するかをプルダウンメニューで選択できます。既定値は「1 日毎」で、「1 日毎」から「30 日毎」までが選択可能です。

- 「毎週、指定された曜日に実行」を選択すると、「日」から「土」までの曜日選択が有効になります。曜日は複数指定できます。指定省略時は「日」曜日のみが選択されます。

インデクシングを繰り返すと、インデックスの中に検索には用いられないデータが蓄積されていきます。これがガベージデータです。ガベージデータが多くなるにつれ検索時間にかかる時間は伸びる傾向があります。(ガベージデータはインデクシングによって発生します。再構築では発生しません)。

以下に述べるガベージコレクションの機能を用いることによって、インデックス内のガベージデータを消去できます。

ガベージコレクション間隔

インデックスのガベージコレクションを実行する間隔、開始時刻、曜日を設定します。インデックス毎に異なった時刻を設定できます。

- 実行間隔を、「設定しない」、「*n* 日毎」、「毎週、指定された曜日に実行」、「*n* 月毎、*n* 日に実行」から選択します。
- 「設定しない」以外を選択すると、「開始時刻 :」で選択した時刻に処理が開始されます。
- 「*n* 日毎」のラジオボタンを選択したときは、何日毎に実行するかをプルダウンメニューで選択できます。既定値は「1 日毎」で、「1 日毎」から「30 日毎」までが選択可能です。
- 「毎週、指定された曜日に実行」を選択すると、「日」から「土」までの曜日選択が有効になります。曜日は複数指定できます。指定省略時は「日」曜日のみが選択されます。
- 「*n* 月毎、*n* 日に実行」のラジオボタンを選択したときは、何か月毎に実行するかと、何日に実行するかを、プルダウンメニューで選択できます。既定値は「1 月毎、1 日に実行」で、月の間隔は 1、2、3、4、6、12 から、日付は 1 日から 31 日までが選択可能です。実行される月は、1 月から数えて *n* 月毎です。(例 : 「4 月毎、15 日に実行」を指定すると、1 月 15 日、5 月 15 日、9 月 15 日の年 3 回の実行)
  ※ある月に存在しない日付が指定されていると、その月には再構築が実行されません。31 日を指定する場合などはご注意ください。29 日 (うるう年以外) や 30 日の指定だと 2 月には処理が実行されません。
- 「インデクシング後に実行」を選択すると、インデクシングの実行の後にガベージコレクションを行います。ある一定以上のガベージデータが存在するときにだけガベージコレクションを実行したいときは、その下の「ガベージデータが全体の *n* パーセント以上のとき実行」または「ガベージデータが全体の *n* 件以上のとき実行」を設定してください。

再構築間隔

インデックスの再構築を実行する間隔、開始時刻、曜日を設定します。インデックス毎に異なった時刻を設定できます。処理対象となるのは、全ファイルです。

- 実行間隔を、「設定しない」、「*n* 日毎」、「毎週、指定された曜日に実行」、「*n* 月毎、*n* 日に実行」から選択します。

- 「設定しない」以外を選択すると、「開始時刻 :」で選択した時刻に処理が開始されます。
- 「*n* 日毎」のラジオボタンを選択したときは、何日毎に実行するかをプルダウンメニューで選択できます。既定値は「1 日毎」で、「1 日毎」から「30 日毎」までが選択可能です。
- 「毎週、指定された曜日に実行」を選択すると、「日」から「土」までの曜日選択が有効になります。曜日は複数指定できます。指定省略時は「日」曜日のみが選択されます。
- 「*n* 月毎、*n* 日に実行」のラジオボタンを選択したときは、何か月毎に実行するかと、何日に実行するかを、プルダウンメニューで選択できます。既定値は「1 月毎、1 日に実行」で、月の間隔は 1、2、3、4、6、12 から、日付は 1 日から 31 日までが選択可能です。実行される月は、1 月から数えて *n* 月毎です。(例 : 「4 月毎、15 日に実行」を指定すると、1 月 15 日、5 月 15 日、9 月 15 日の年 3 回の実行)
  ※ある月に存在しない日付が指定されていると、その月には再構築が実行されません。31 日を指定する場合などはご注意ください。29 日 (うるう年以外) や 30 日の指定だと 2 月には処理が実行されません。

検索結果設定

検索結果画面に表示される、検索で見つかったファイルの一部の表示形態を選択します。

「ファイルの先頭を表示」では常にファイルの先頭を表示します。「検索語の周辺文章を表示」では、検索語が検索画面に現れるようにするために、ファイルの文章中から検索語の周辺を切り出して表示します。
検索結果設定の変更は、再構築実行後に有効になります。

使用する類語辞書セット

インデクシングおよび検索のときに使用される類語辞書セットを指定します。
類語機能を使用しない場合は、初期設定の通り、類語辞書セットを「なし」のままの設定にします。

インデクシング、ガベージコレクション、再構築「手動実行」ボタン

左側の欄に存在する「手動実行」ボタンを押すと、すぐにインデクシング、ガベージコレクションまたは再構築が開始されます。

オプションパックの導入や辞書の追加や変更をしたときなど、インデックスを作り直す必要のあるときは、「再構築間隔」の欄の方にある「手動実行」ボタンをクリックして、インデックスを再構築します。

注意 1

インデクシングの処理実行中に、インデックス追加や削除の操作はしないでください。このような操作をすると、以後、インデクシングや検索が正しく動作しなくなることがあります。

注意 2

Red Hat Enterprise 4.0 **を使用している場合は、最初のインデックスを新規追加した直後に** crond **を再起動しないと、間隔の設定が有効になりませんので注意してください。**

使用メモリサイズの上限値

この値を大きくすると、インデクシングにかかる時間が短縮されることがあります。ただし、ホストの物理メモリが少ない場合などは、かえって遅くなることもあるため注意が必要です。

ファイルサイズの上限値

インデクシングの対象となるファイルの大きさです。この値よりも大きなファイルはインデクシングの対象にはなりません。意図しない検索漏れが発生しないように注意して設定してください。

テキストの上限値

インデクシングの対象となるファイルのテキスト部分の大きさです。この値よりも大きなテキストのサイズのファイルはインデクシングされません。

単語長の上限値

インデクシングの対象となる単語の長さです。この値よりも長い単語はインデクシングされません。この場合、インデクシングされないのは単語のみで、その単語を含むファイルはインデクシングされます。

注意 3

「使用メモリサイズの上限値」、「ファイルサイズの上限値」、「テキストの上限値」、「単語長の上限値」といった上限値を増やすにあたっては、それに応じたハードウェア性能が、正常動作のためには必要です。

インデクシングに際しての注意

Kabayaki Enterprise は、1 つのインデックスに対して、同時に複数のインデクシングプロセスを実行することを禁止しています。そのため、インデクシング実行中に同じインデックスに対してインデクシングを実行しようとすると、以下のようなメッセージがログに出力されることがあります。

> "... not executed, because /*vardir*/kabayaki/targets/*idxname*/lock exists."

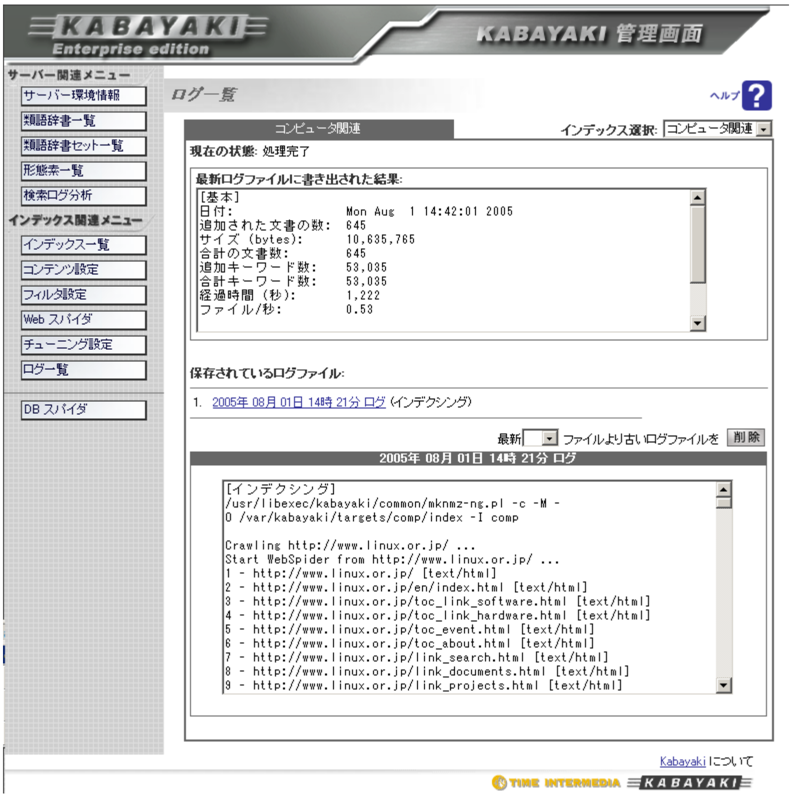
vardir には、Red Hat なら /var が入ります。*idxname* に入るのはインデックスの内部名です。

外部的または内部的な要因によって Kabayaki Enterprise のプロセスが意図しない異常終了を起こしたときも、該当するインデックスのログ一覧には上記のようなメッセージが表示されます。このような場合は、ロックファイルがファイルシステムに残ったままになっていることにより、それ以後のインデクシングができなくなることがあります。この状態から再びインデクシング実行を可能にするためには、以下の手順で作業してください。

1. 管理画面の左側に表示されるメニューの「ログ一覧」ボタンをクリックして、ログ一覧を表示させます。
(ログについては、次の節の「ログ一覧」を参照してください)。
2. 「ログ一覧」で上記のメッセージを確認して、ロックファイル /*vardir*/kabayaki/targets/*idxname*/lock を手動で削除します。
3. /*vardir*/kabayaki/index/*idxname*/NMZ.lock2 ファイルが存在していたならば、そちらも削除します。
4. 「チューニングの画面」で、インデクシングまたは再構築の手動実行ボタンを押して、処理が正常に開始されるかどうかを確認します。

ログ一覧

インデクシングやガベージコレクションに関するログを表示します。ホストのシステム管理者は、ログを定期的に参照することによって、処理状況を確認できます。管理画面の左側に表示されるメニューの「ログ一覧」ボタンをクリックすると、次のような画面が表示されます。



現在の状態：

実行中の処理に応じて

- 「インデクシング処理中」
- 「ガベージコレクション処理中」
- 「インデクシング / ガベージコレクション処理中」

と表示されます。処理が完了していると

- 「処理完了」

と表示されます。「処理完了」のときはその下に表示される「最新ログファイルに書き出された結果 :」で、処理結果を参照できます。

保存されているログファイル：

インデクシングやガベージコレクションの実行結果のログファイルが、日付の新しいものから一覧表示されます。日付の部分をクリックすると、画面下部のテキスト領域に詳細なログが表示されます。ガベージコレクションをインデクシング後に設定した場合、一つのログに表示されます。

詳細表示される項目には次の情報があります。

インデクシング

- 日付
- 追加された文書数
- 削除された文書数
- サイズ
- 更新された文書数
- 合計の文書数
- 追加キーワード数
- 合計キーワード数
- わかち書き・経過時間
- ファイル / 秒・システム
- Namazu
- エラー、警告、詳細

ガベージコレクション (手動実行の場合)

- ガベージデータ数
- ガベージの割合

ガベージコレクション (手動実行以外の場合)

- 設定されていたガベージデータ数
- 設定されていたガベージの割合
- ガベージデータ数
- ガベージの割合

最新 *n* ファイルより古いログファイルを削除

古いログファイルを自動的に削除することが可能です。初期設定ではログは自動的に削除されません。

プルダウンメニューから残したいログの数を選択して「削除」ボタンを押すと、その時点で指定された数のログファイルだけ残し、古いログファイルが削除されます。また、それ以降のインデクシングの度に古いファイルから順に削除され、常に指定された数のログファイルだけが残るようになります。

数値指定なし (空白) の場合はログの個数を制限しないため、自動削除は実行されません。

第 4 章　検索および検索結果画面

検索画面の表示

インデックス作成後に実際に検索するには、Microsoft Internet Explorer や Netscape Communications Netscape Navigator などの Web ブラウザに以下のような URL を入力して、検索画面を表示させます。

http:// *ホスト名* /kabayaki/

「*ホスト名*」の部分には、Kabayaki Enterprise をインストールしたコンピュータの名前を入力します。

たとえば、Kabayaki Enterprise をインストールしたホストが search.timedia.co.jp ならば、次の URL になります。

http://search.timedia.co.jp/kabayaki/

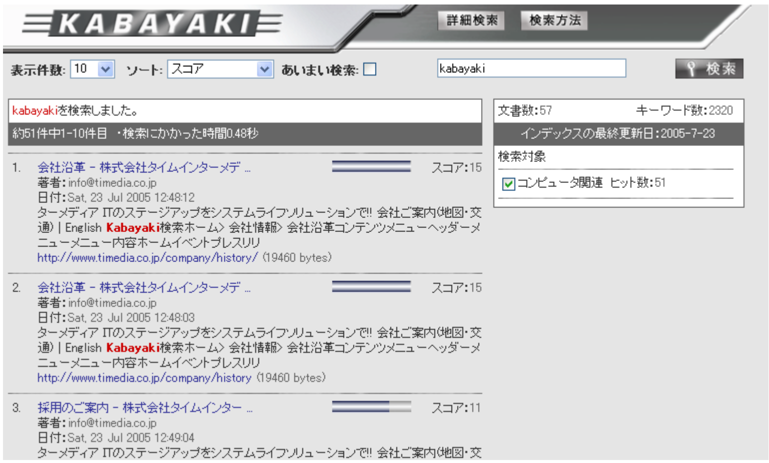
Web ブラウザには、次のような検索画面が表示されます。

検索画面 :

「表示件数」には検索結果画面に一度に表示する件数、「ソート」には検索結果画面上での並べ替え順を指定します。また、「あいまい検索」のチェックボックスで、類語機能を使用した検索を実行するかどうかを指定します。

検索文字列を指定して「検索」ボタンを押すと、検索結果画面が表示されます。

検索結果画面 :



検索方法

検索式

単一単語検索

調べたい単語を 1 つ指定するだけのもっとも基本的な検索手法です。
例 :

namazu

AND 検索

ある単語とある単語の両方を含む文書を検索します。検索結果を絞り込むのに有効です。3 つ以上の単語を指定することも可能です。単語と単語の間に and または & を挿みます。
例 :

Linux and Netscape

and または & は省略できます。単語を空白で区切って羅列するとそれらの語すべてを含む文書を AND 検索します。

OR 検索

ある単語とある単語のどちらかを含む文書を検索します。3 つ以上の単語を指定することも可能です。単語と単語の間に or または | を挿みます。
例 :

Linux or FreeBSD

NOT 検索

ある単語を含み、ある単語を含まない文書を検索します。3 つ以上の単語を指定することも可能です。単語と単語の間に not または ! を挿みます。
例 :

Linux not UNIX

グループ化

AND 検索、OR 検索、NOT 検索を括弧でグループ化できます。括弧の両隣には空白を入れる必要があります。
例 :

(Linux or FreeBSD) and Netscape not Windows

部分一致検索

部分一致検索には前方一致、中間一致、後方一致の 3 種類があります。

- 前方一致検索
  inter* (inter から始まる単語を含む文書を検索)
- 中間一致検索
  text (text を内包する単語を含む文書を検索)
- 後方一致検索
  *net (net で終わる単語を含む文書を検索)

フィールド指定の検索

Subject:、‘ From:、Message-Id: といったフィールドを指定して検索する手法です。特に Mail/News のファイルを扱う際に効果を発揮します。例 :

- +subject:Linux
  (Subject: に Linux が含まれる文書)
- +subject:"GNU Emacs"
  (Subject: に GNU Emacs が含まれる文書)
- +from:foo@bar.jp
  (From: に foo@bar.jp が含まれる文書)
- +message-id:<199801240555.OAA18737@foo.bar.jp>
  (Message-Id を指定)

特記事項

- いずれの検索方法でもアルファベットの大文字・小文字の区別 はしません。
- 日本語の複合語は形態素単位に分割し、それらをフレイズ検索します。分割は不適切に行なわれることがあります。
- JIS X 0208 (いわゆる全角文字) の英数字と記号の一部 (ASCII と重複しているもの) は ASCII (いわゆる半角文字) として処理されます。
- 記号を含む語の検索ができます。例 : TCP/IP。ただし、記号の処理は完全ではないので TCP and IP のように分割して AND 検索をかけた方が取りこぼしがありません (その代わり余計なファイルまでヒットしてしまう可能性があります)。

- 中間一致・後方一致、正規表現、フィールド指定の検索には少し時間がかかります。
- and, or, not を単語として検索したいときはそれぞれ、"..." と 2 重引用符で、あるいは {...} と中括弧で囲みます。

ソート

標準の Kabayaki Enterprise パッケージでは、検索結果を、スコア、日付、題名、著者、サイズ、URI の順で並べ替えることができます。既定値はスコア順でのソートです。

スコア

- 基本的には、指定された検索文字列の登場回数が多い文書ほどスコアの数値は高くなります。
- 検索文字列が文書のどの場所に含まれるかで、スコアに加算される数値が変わることがあります。登場回数が同じ 1 回でも、たとえば検索文字列がタイトルに使用されている文書は、そうでない文書よりもスコアが高くなります。

複数の検索文字列が指定された場合は、ヒットした文書数の少ない文字列をより多く含む文書の方がスコアが高くなる傾向があります。

付録 A　Kabayaki Enterprise 文書フィルタとプロパティ検索詳細

文書フィルタ

Kabayaki Enterprise は、文書フィルタと呼ばれる外部プログラムと連携して、HTML やテキストファイル以外の文書形式も検索の対象とすることができます。Kabayaki Enterprise に標準で添付されている文書フィルタは、Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF について、日本語全文検索システム Namazu の文書フィルタと比べ、より高い精度の検索を提供します。

ジャストシステム 一太郎、富士通 OASYS 等、Namazu では検索できない文書形式にも対応します。

対応文書

- テキスト文書 JIS/EUC/SJIS
- HTML 文書
- Microsoft Word 95/97/98/2000/2002(XP)/2003
- Microsoft Excel 95/97/2000/2002(XP)/2003
- Microsoft PowerPoint 95/97/2000/2002(XP)/2003
- Microsoft Rich Text Format
- ジャストシステム 一太郎 5 〜 13
- 富士通 OASYS V6/V7
- Adobe Portable Document Format Acrobat 4.0/5.0/6.0(PDF 1.3/1.4/1.5)
- Mail/News

その他の文書については、namazu-2.0.14 と同等です。

プロパティ検索

プロパティとは、文書を識別するためにファイルに埋め込まれた文書の属性情報のことです。このプロパティを検索の対象とすることができます。プロパティには、タイトル、作成者の名前、キーワード、コメントなどを記録されています。プロパティに関しては、各文書を作成するアプリケーションの取扱説明書をご覧ください。

表 1 Kabayaki Enterprise で検索できるプロパティ

プロパティ	説明	**Word**	**Excel**	**PowerP**	**PDF**	**一太郎**	**OASYS**
title	タイトル	○	○	○	○	×	×
author	著者	○	○	○	○	×	×
company	会社名	○	×	○	×	×	×

OLE オブジェクト検索

Kabayaki Enterprise では、OLE で埋め込まれたオブジェクトも検索の対象とすることができます。OLE に関しては、OLE をサポートしているアプリケーションの取扱説明書をご覧ください。

表 2 Kabayaki Enterprise で検索できる OLE オブジェクト

Dest/Source	**Word**	**Excel**	**PowerPoint**
Word	○	○	×
Excel	○	○	×
PowerPoint	×	×	×